Présentation et évaluation d'un modèle d'attention audiovisuelle sur une base de scènes de conversations dynamiques

Antoine **COUTROT**[a], Nathalie **GUYADER**[a]
[a] *Gipsa-lab, CNRS & Université de Grenoble-Alpes, France*

ABSTRACT

Les modèles de saillance classiques ne prennent pas en compte les aspects "sociaux" de la perception visuelle, et donnent des résultats peu satisfaisants dès que des visages sont présents à l'écran. En effet ces modèles considèrent rarement les visages, et jamais l'information sonore, pourtant critique pour modéliser l'attention visuelle dans des scènes de conversations entre plusieurs personnes. Dans cette étude, nous proposons un modèle de saillance audiovisuelle pour prédire les régions les plus susceptibles d'attirer les regards de personnes explorant librement des scènes dynamiques de conversations. Nous montrons que notre modèle présente de meilleurs résultats que de précédents ne prenant pas en compte l'information sonore et attribuant une valeur de saillance égale et constante à tous les visages.

INTRODUCTION

Le but des modèles d'attention visuelle est de prédire les régions d'une scène qui seront le plus probablement fixées lors d'une exploration libre. La plupart des modèles précédemment développés séparent la scène observée en plusieurs cartes d'attributs visuels statiques (contraste, contours...) et/ou dynamiques (mouvement) [1]. Pour chacune des cartes, les régions contrastant le plus avec leur voisinage sont mises en valeur, puis fusionnées au sein d'une carte dite "de saillance maitresse". Cependant, ces modèles ne prennent pas en compte de nombreux aspects de la perception visuelle, et ont de faibles performances dès qu'il s'agit de modéliser l'attention visuelle pour des scènes à fort contenu sémantique [2]. Les scènes de conversation font partie de ces dernières. En effet, alors que les visages attirent l'attention de manière particulièrement efficace [3], les modèles de saillance "classiques" ne les prennent pas en compte [4]. Pour pallier à ce manque, certains auteurs ont utilisé des détecteurs de visages afin d'obtenir des cartes de saillance « visage ». L'ajout de ces cartes aux autres attributs augmente considérablement les performances de leurs modèles [5]. Jusqu'à présent, seuls les attributs visuels ont été exploités, et les informations issues d'autres modalités, comme l'audition, ne sont que très rarement considérées. Cependant, notre perception du monde est multimodale, et de nombreuses études ont établi que le son joue un rôle important dans la perception et l'exploration de scènes [6]. Dans de précédents travaux, nous avons montré que si le son avait un effet limité sur les mouvements oculaires enregistrés lors de l'exploration libre de scènes dynamiques au contenu varié (objets, paysages...) [7], il devenait un facteur essentiel pour expliquer les mouvements oculaires lors de l'exploration d'une conversation entre plusieurs personnes [8]. Afin de quantifier l'importance relative des visages et des autres attributs visuels pour expliquer des mouvements oculaires nous avons proposé un modèle statistique [9]. Ce modèle nous a permis de montrer que les attributs de bas niveau (saillance statique, dynamique, biais de centralité) attirent le regard bien moins efficacement que les visages, et particulièrement que les visages des locuteurs. Dans l'étude que nous proposons ici, nous utilisons les résultats issus de cette quantification afin construire un modèle de saillance audiovisuelle pour des scènes de conversations. Contrairement aux précédents modèles donnant un poids égal et constant à tous les visages [5], nous utilisons ici un algorithme d'identification du locuteur (*speaker diarization*) prenant en compte l'information issue de l'image et de la bande-son afin de déterminer "qui parle quand", et de donner un poids supérieur à la saillance du locuteur par rapport à celle des auditeurs. Afin d'évaluer les performances de notre modèle de saillance audiovisuelle, nous avons mené une expérience d'oculométrie. Nous avons enregistré les mouvements oculaires de 40 personnes regardant 15 vidéos, dans lesquelles 4 personnes discutent lors d'une réunion de travail. Ces vidéos sont issues d'une base disponible sur internet [10]. Chaque vidéo a été vue par 20 personnes avec sa bande-son originale, et par 20 autres personnes sans son. En comparant les régions regardées par les participants aux cartes de saillance, nous montrons que notre modèle est meilleur que ceux qui ne prennent pas en compte l'information sonore.

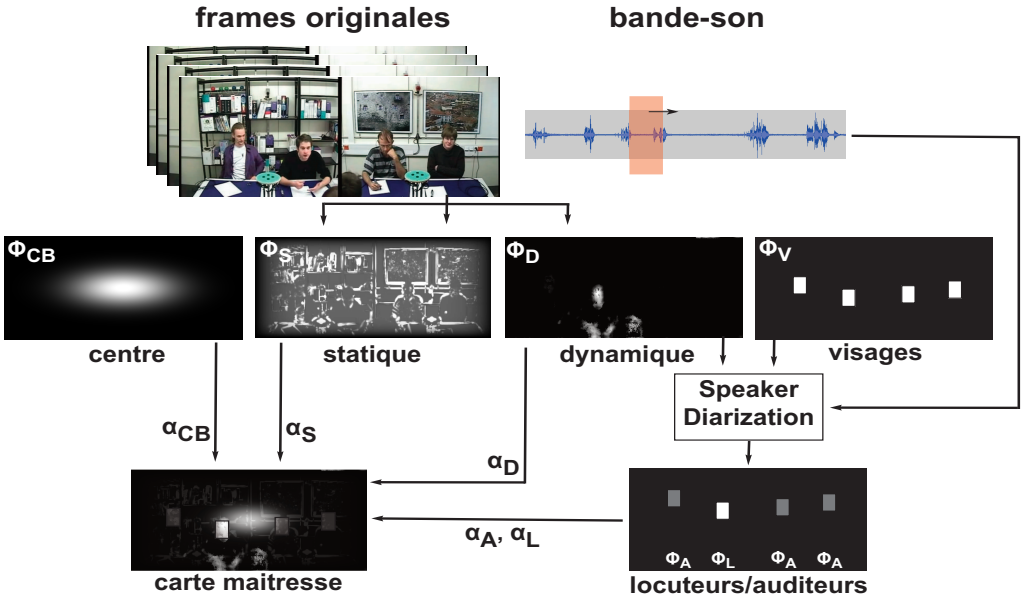

Figure 1 - **Diagramme du modèle de saillance audiovisuelle. De chaque frame sont extraits une carte correspondant au biais de centralité, à la saillance statique, à la saillance dynamique, et aux visages présents à l'image. Un algorithme de "speaker diarization" utilisant l'information de la bande-son permet de séparer les auditeurs des locuteurs. Un poids plus important est donné à ces derniers lors de la fusion donnant la carte de saillance maitresse.**

MODÈLE DE SAILLANCE AUDIOVISUELLE

Architecture

Chaque vidéo est d'abord décomposée en quatre cartes d'attributs visuels connus pour leur rôle dans l'allocation attentionnelle, comme l'illustre la Figure 1.

- **Biais de centralité** de nombreuses études oculométriques ont montré que les participants ont davantage tendance à fixer le centre que les bords des stimuli visuels. Plusieurs hypothèses ont été proposées pour expliquer ce biais. Certaines sont liées aux stimuli, comme le biais du photographe (les régions d'intérêt sont plus souvent au centre de l'image), d'autres sont inhérentes au système oculomoteur (la position de repos des yeux est le centre) ou aux stratégies d'exploration (le centre est un endroit stratégique pour observer une scène) [13]. Ici, nous avons modélisé le biais de centralité par une gaussienne centrée au milieu de chaque frame **(Φ_D)**.

- **Saillance bas niveau** la saillance bas niveau de chaque frame vidéo est calculée grâce à un modèle de saillance visuelle bio-inspiré [11]. Ce modèle, basé sur l'information de luminance, décompose la saillance de chaque frame en une carte **statique (Φ_S)** et une carte **dynamique (Φ_D)**. La voie statique utilise les hautes fréquences spatiales pour repérer les zones de fort contraste. La voie dynamique compense dans un premier temps le mouvement de la caméra pour ne détecter que le mouvement par rapport l'arrière-plan. Ensuite, les composantes basse fréquence de deux frames successives sont utilisées pour extraire le mouvement.

- **Visages** les visages de chaque personne présente à l'image sont repérés grâce à un programme permettant la segmentation semi-automatique des objets dans des vidéos **(Φ_V)** [12]. Les locuteurs **(Φ_L)** ont été séparés des auditeurs **(Φ_A)** via un algorithme de "Speaker Diarization" permettant de repérer "qui parle quand". Cet algorithme utilise l'information issue de la bande-son, des cartes de saillance dynamique, statique, des personnes et a été décrit en détail en [15].

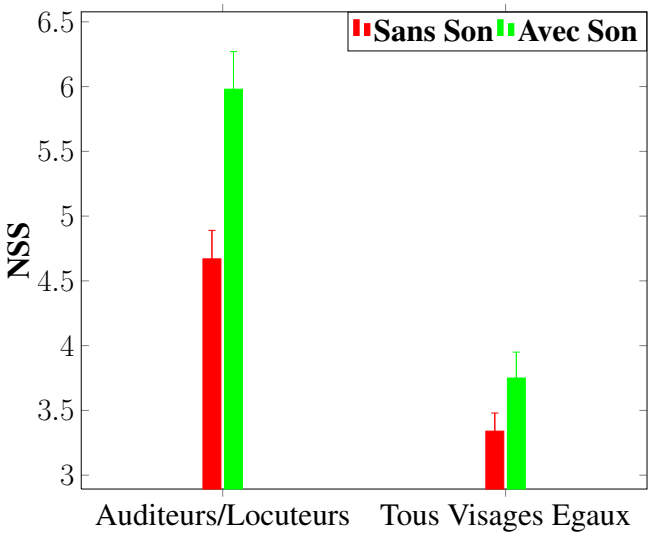

Figure 2 - Normalized Scanpath Saliency de deux modèles : (1) les poids du visages des locuteurs et des auditeurs sont estimés séparément, (2) tous les visages ont le même poids. Les NSS sont calculées dans les deux conditions sonores.

Ces cartes sont calculées pour chaque frame, pondérées et combinées linéairement en une carte de saillance maitresse M : $M(f) = \sum_{k \in \{CB,S,D,A,L\}} \alpha_k(f) \Phi_k(x,y,f)$

Les poids $\boldsymbol{\alpha_k}$ été optimisés grâce au Lasso, une méthode statistique permettant d'estimer, pour un jeu d'observation donné, l'importance relative de plusieurs variables. Ici, les observations sont les positions oculaires enregistrées au cours de l'expérience décrite ci-dessous. Le Lasso retourne donc un poids pour chaque attribut et chaque frame. Cette estimation montre que dans les scènes de conversation, les attributs les plus important pour prédire les positions oculaires sont les visages, et plus particulièrement le visage des locuteurs (voir [9] pour plus de détails). Pour chaque attribut k, nous avons moyenné ses poids $\boldsymbol{\alpha_k}$ sur l'ensemble des frames de chaque vidéo.

EVALUATION

Comparer les zones prédites comme étant saillantes avec les positions oculaires de participants à une expérience oculométrique est une manière classique d'évaluer un modèle de saillance. Nous avons demandé à 40 personnes de regarder 15 vidéos (1232 x 504 pixels) durant entre 19 s et 1 min 20 s. Les stimuli sont des vidéos représentant 4 personnes durant une réunion de travail et font partie du corpus AMI, librement disponible sur internet [10]. Les bandes-son sont monophoniques, échantillonnées à 48 kHz, et les dialogues sont en anglais. Les participants ont vu la moitié des stimuli avec leur bande-son originale et l'autre moitié sans aucun son. Au final, chaque stimuli a été vu par 20 différents participants dans chaque condition sonore. Pour évaluer notre modèle, nous avons utilisé une métrique largement répandue dans la littérature, le Normalized Saliency Scanpath (NSS) [16].

$$NSS = \frac{M_m M_p - mean(M_m)}{std(M_m)}$$

Le NSS correspond aux valeurs centrées réduites que la carte de saillance prédite par le modèle (M_m) prend aux positions oculaires des participants, consignées sur une carte de positions oculaires (M_p). Plus le NSS est grand, meilleure est la performance du modèle de saillance. Nous avons calculé le NSS de deux modèles basés sur l'architecture présentée ci-dessus (1) en différentiant les locuteurs des auditeurs et (2) sans différentier les locuteurs des auditeurs, comme c'est jusqu'à présent le cas dans la littérature [5]. Nous comparons également les NSS selon les conditions sonores (voir Figure 2). Pour ne pas évaluer le modèle

avec les mêmes positions oculaires que celles qui nous ont servi à le construire, nous avons utilisé la méthode du "*leave one out*", c'est-à-dire les poids des attributs d'une vidéo donnée sont issus de la moyenne des poids de toutes les vidéos, sauf de celle traitée. Nous observons que les performances du premier modèle sont nettement supérieures à celles du second, et que les positions oculaires enregistrées avec la bande-son originale sont mieux prédites que celles enregistrées sans aucun son (tests de Student pairés, tous les $p<.001$).

CONCLUSION

Dans cette étude, nous avons proposé un modèle de saillance audiovisuelle permettant de prédire les positions oculaires sur une scène de conversation. Notre modèle s'appuie à la fois sur des attributs visuels bas niveau (saillance statique, saillance dynamique, biais de centralité), haut niveau (visages) et sur des attributs auditifs (parole). Afin d'estimer les poids relatifs de chacun de ces attributs, nous utilisons une méthode statistique (Lasso). Les locuteurs sont différenciés des auditeurs au moyen d'un algorithme de *Speaker Diarization*. Cette méthode et cet algorithme ont tout deux été présentés dans de précédentes études [9,15]. Ce modèle a été évalué sur une base de vidéos de conversations standardisées, grâce à une expérience oculométrique. Estimer séparément les poids du visage des locuteurs et des auditeurs permet de considérablement améliorer les performances du modèle, et supprimer l'information sonore rend les positions oculaires moins prédictibles. De futurs travaux pourront estimer le poids d'autres parties du corps que le visage (mains, torse), pour enrichir les cartes non plus visages mais "personnes" de ce modèle.

RÉFÉRENCES

[1] A. Borji and L. Itti, “State-of-the-art in Visual Attention Modeling”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

[2] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, “Eye guidance in natural vision: Reinterpreting salience,” *Journal of Vision*, vol. 11, no. 5, pp. 1–23, May 2011.

[3] S. M. Crouzet, H. Kirchner, and S. J. Thorpe, “Fast saccades toward faces: Face detection in just 100 ms,” *Journal of Vision*, vol. 10, no. 4, pp. 1–17, Apr. 2010.

[4] E. Birmingham, W. F. Bischof, and A. Kingstone, “Saliency does not account for fixations to eyes within social scenes,” *Vision Research*, vol. 49, pp. 2992–3000, 2009.

[5] S. Marat, A. Rahman, D. Pellerin, N. Guyader, and D. Houzet, “Improving Visual Saliency by Adding ‘Face Feature Map’ and “Center Bias”", *Cognitive Computation*, vol. 5, no. 1, pp. 63–75, 2013.

[6] S. Onat, K. Libertus, and P. König, “Integrating audiovisual information for the control of overt attention”, *Journal of Vision*, vol. 7, no. 10, pp. 1–16, 2007.

[7] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, “Influence of soundtrack on eye movements during video exploration”, *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[8] A. Coutrot and N. Guyader, “Toward the Introduction of Auditory Information in Dynamic Visual Attention Models,” presented at the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS 2013), Paris, France, 2013.

[9] A. Coutrot and N. Guyader, “How Saliency, Faces and Sound influence gaze in Dynamic Social Scenes”, *Journal of Vision*, in press.

[10] Carletta, J. (2006), "Announcing the AMI Meeting Corpus", The ELRA Newsletter 11(1), p. 3-5.

[11] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, “Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos,” *International Journal of Computer Vision*, vol. 82, no. 3, p. 231–243, 2009.

[12] P. Bertolino, “Sensarea: an Authoring Tool to Create Accurate Clickable Videos,” presented at the 10th Workshop on Content-Based Multimedia Indexing, Annecy, France, 2012, pp. 1–4.

[13] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, “Quantifying center bias of observers in free viewing of dynamic natural scenes,” *Journal of Vision*, vol. 9, no. 7, pp. 1–16, Jul. 2009.

[15] A. Coutrot, and N. Guyader, “An Audiovisual Attention Model for Natural Conversation Scenes,” *IEEE International Conference on Image Processing*, Paris, France, 2014.

[16] O. Le Meur and T. Baccino, “Methods for comparing scanpaths and saliency maps: strengths and weaknesses,” *Behavior Research Methods*, vol. 45, no. 1, pp. 251–266, 2013.